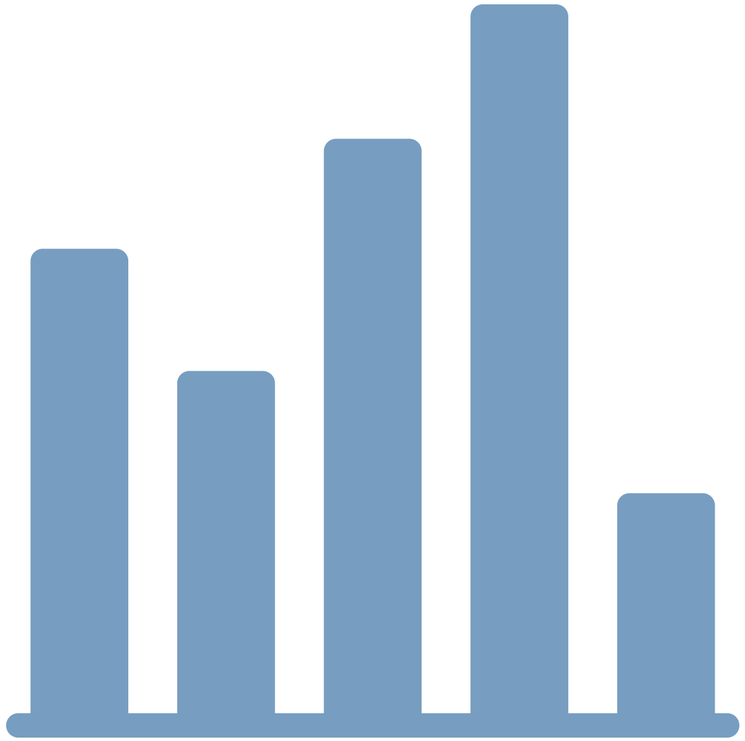
University of Washington
Holli Meyers, Yuki Sherard, Britteny Dressen

From Data to Destination

ML Bike Rental Predictions

Understanding the Problem



Considerations

Capital Bike Sharing (CBS) aims to **stay competitive** in the growing transportation industry, particularly in their local area of Washington, D.C.

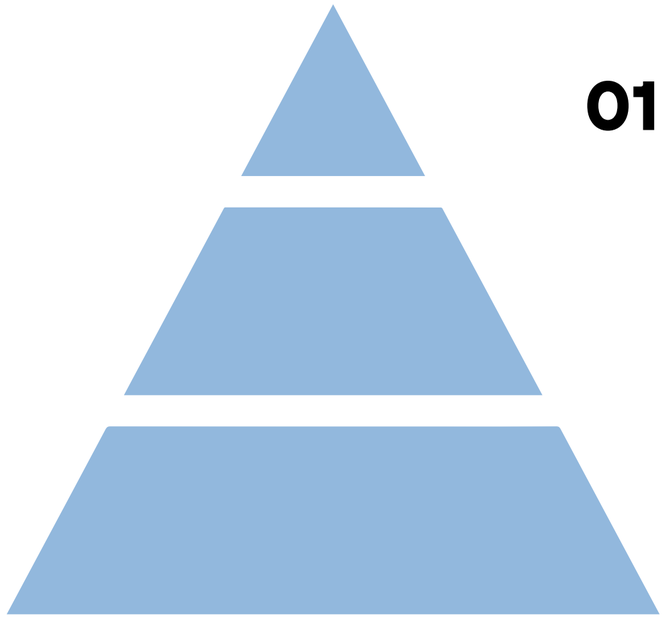
An oversupply of bikes can clutter urban spaces, reduce **efficiency**, and strain city infrastructure.

Accurately forecasting rental bike demand is crucial for **ensuring availability**, minimizing waste, and enhancing user satisfaction.

Questions

How accurately can we predict bike rental demand to ensure optimal fleet availability without oversupply or shortages?

At what point does bike inventory exceed demand, and how can we identify surplus thresholds to guide removal or reallocation decisions?



01

02

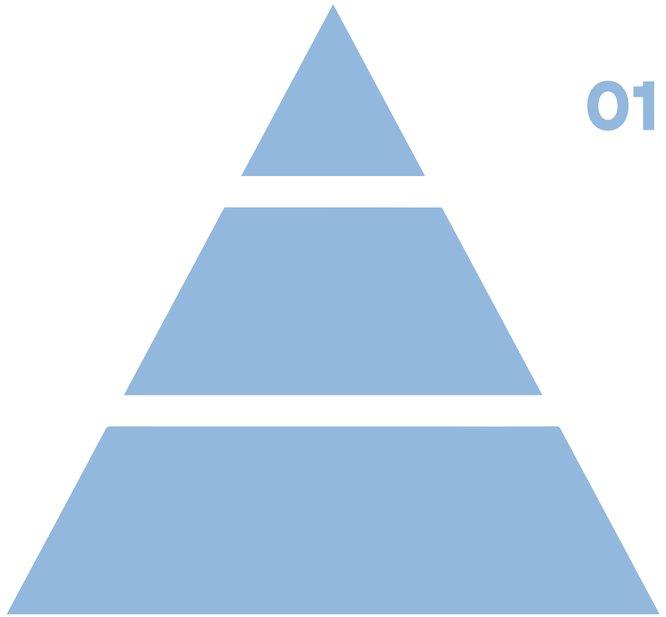
03

Analysis Workflow

EDA

Visualize and summarize the dataset to gather an overview of its characteristics (i.e., structure, key variables, and trends).

Then, extract insights to guide modeling—identify patterns, outliers, seasonality, and correlations between features and purchase volume.



01

02

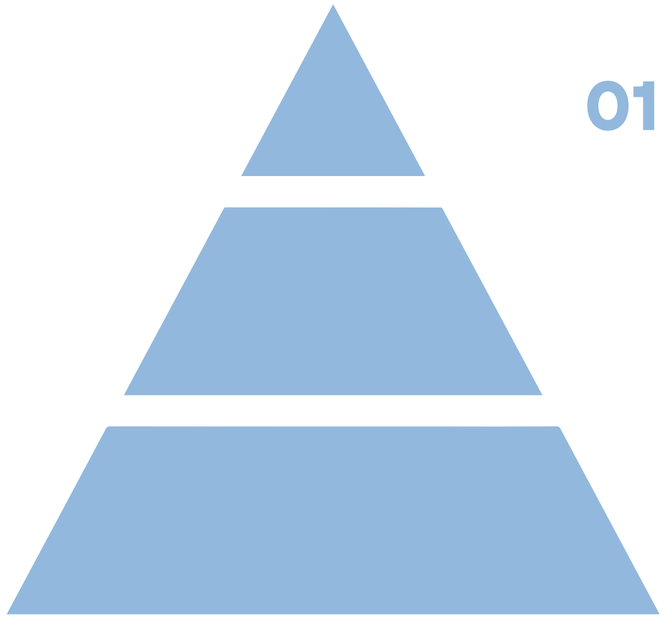
03

Analysis Workflow

Data Preparation

Clean and preprocess the dataset by handling missing values, formatting timestamp variables, and encoding categorical features.

Engineer new features to better capture purchase behavior patterns.



Analysis Workflow

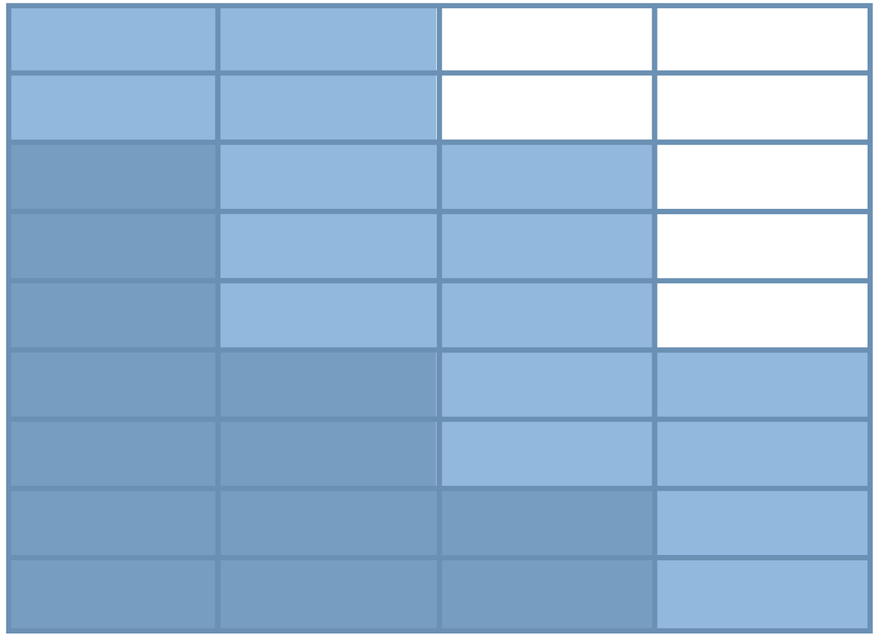
Modeling

Train and compare multiple machine learning models to predict bike rental demand.

Iterate on model selection, feature sets, and hyperparameters based on validation metrics, with the goal of balancing accuracy and interpretability.

About the Dataset

Contains two years of data from 2011 to 2012.



Season

Timestamps

Weather conditions

User types (casual or registered)

Total number of rented bikes

Fanaee-T,Hadi. (2013). Bike Sharing Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.

Data Wrangling

Normalization

Temperature, wind speed, and humidity variables came standardized.

Models were tested using normalized and original values to assess impact on performance.

One-Hot Encoding

Binary columns were converted to factors for compatibility with modeling techniques.

Applied a model matrix transformation to categorical variables like weekday, season, and weather condition.

Multiple model variants were tested with and without one-hot encoded features.

Feature Selection

Removed redundant columns such as casual and registered rider counts to avoid data leakage—they sum to the target variable, total rentals.

Columns like year, date, and rental ID were excluded from modeling due to lack of predictive value or uniqueness.

Dashboard Overview

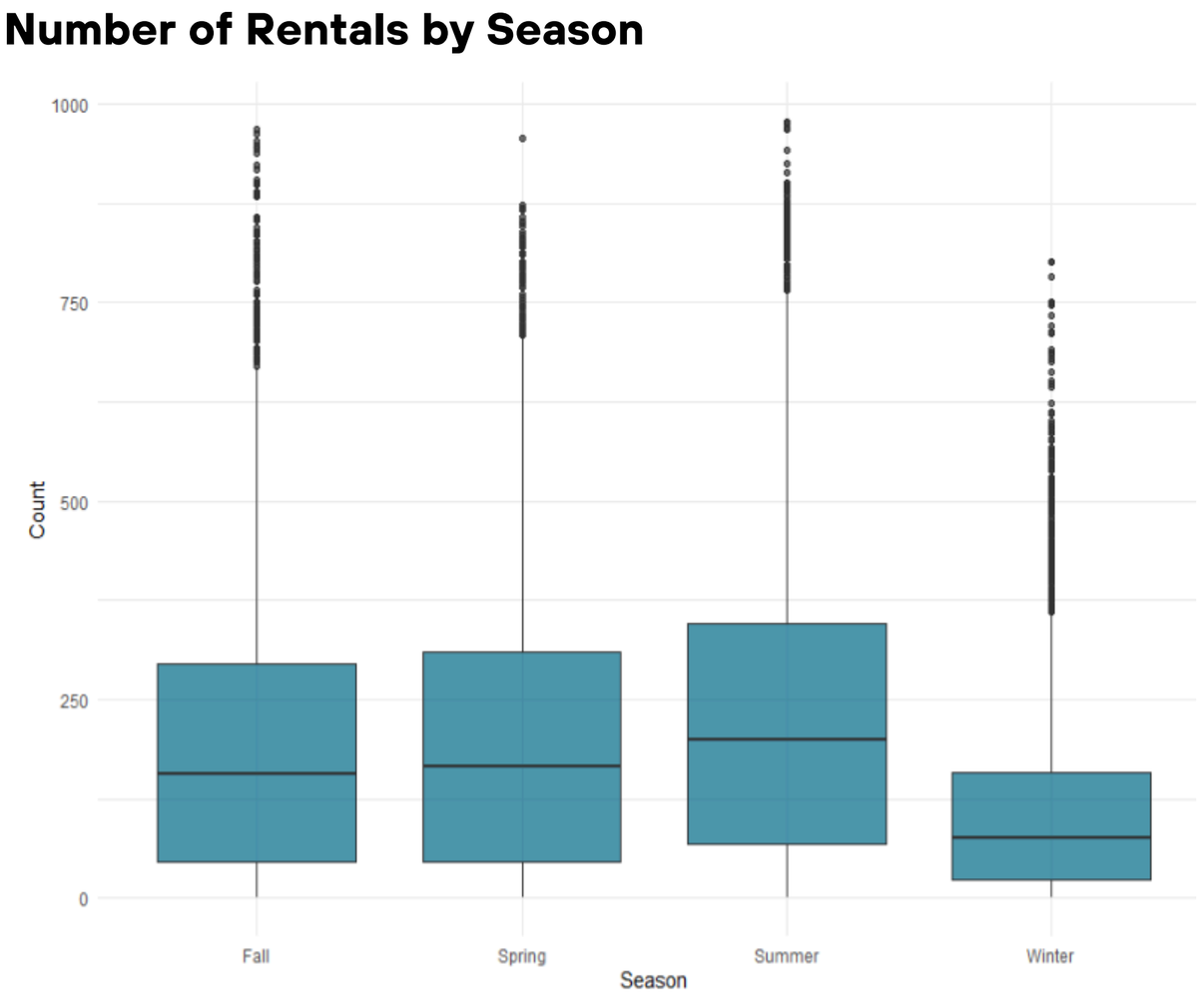
Machine learning model dashboard in R with Flexdashboard.

EDA

Hypothesis

Rental demand increases under the following conditions:

- Warmer temperatures and favorable weather (e.g., clear or partly cloudy conditions).
- Weekends and holidays, due to increased leisure activity.
- Weekday mornings and afternoons, reflecting commuter patterns.

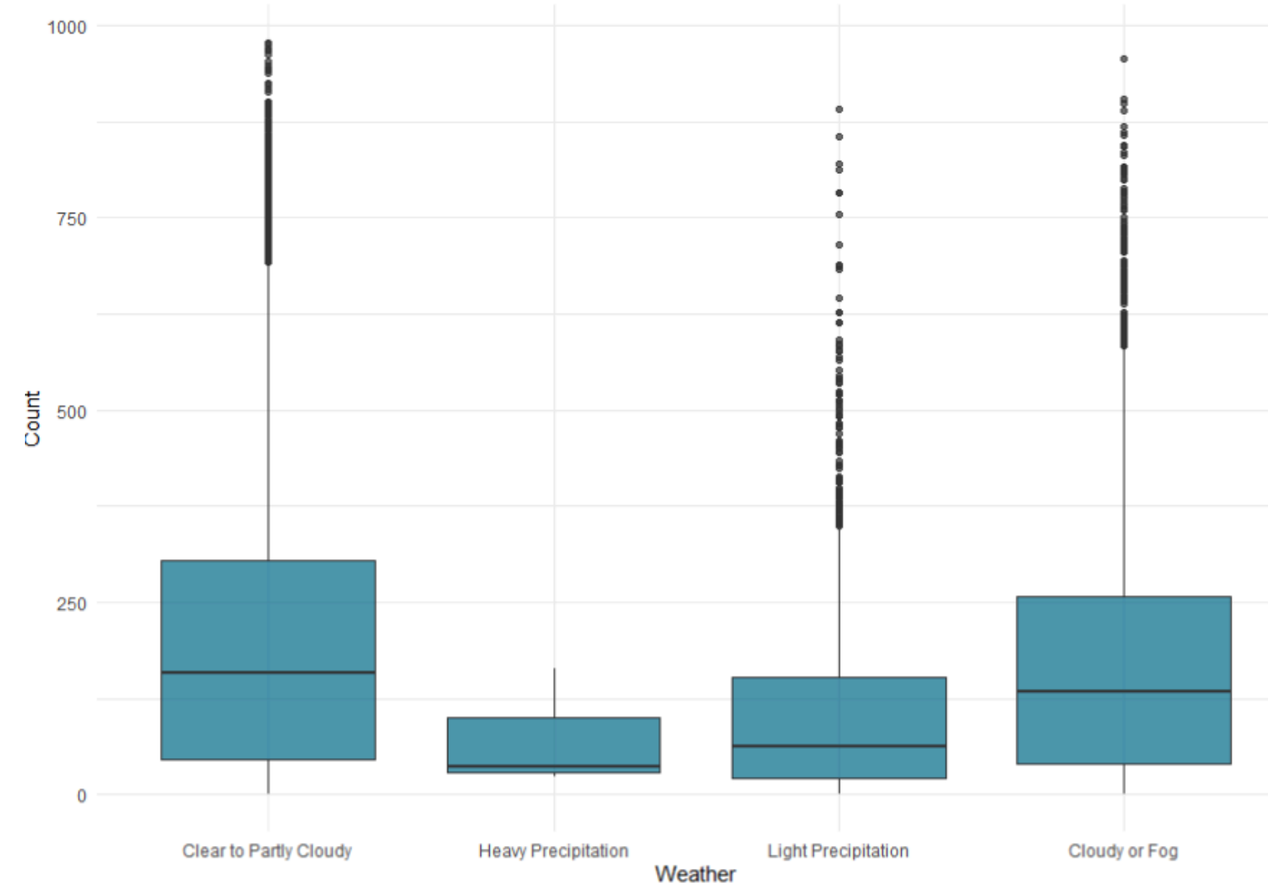


EDA

Hypothesis

Rental demand increases under the following conditions:

- Warmer temperatures and favorable weather (e.g., clear or partly cloudy conditions).
- Weekends and holidays, due to increased leisure activity.
- Weekday mornings and afternoons, reflecting commuter patterns.



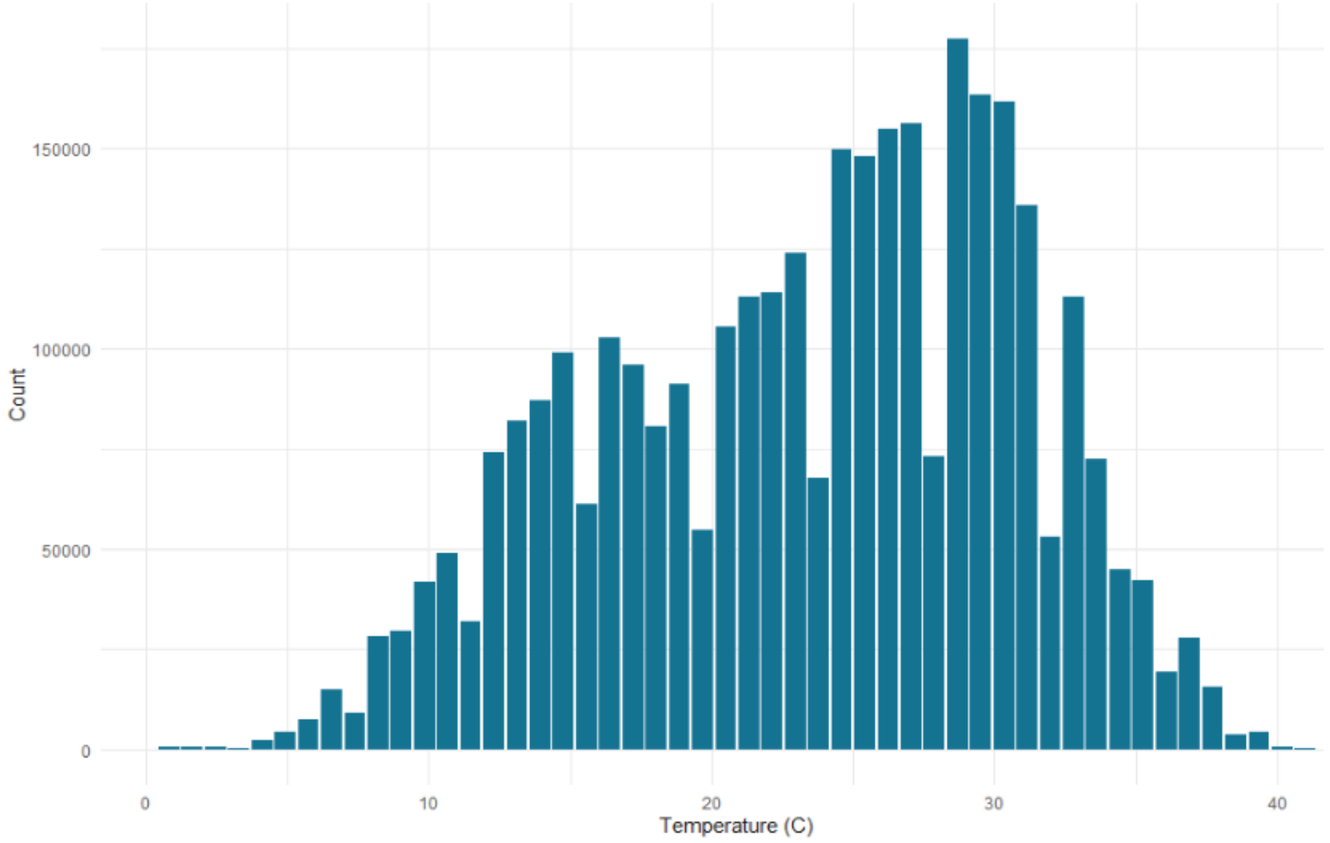
EDA

Hypothesis

Rental demand increases under the following conditions:

- Warmer temperatures and favorable weather (e.g., clear or partly cloudy conditions).
- Weekends and holidays, due to increased leisure activity.
- Weekday mornings and afternoons, reflecting commuter patterns.

Number of Rentals by Temperature

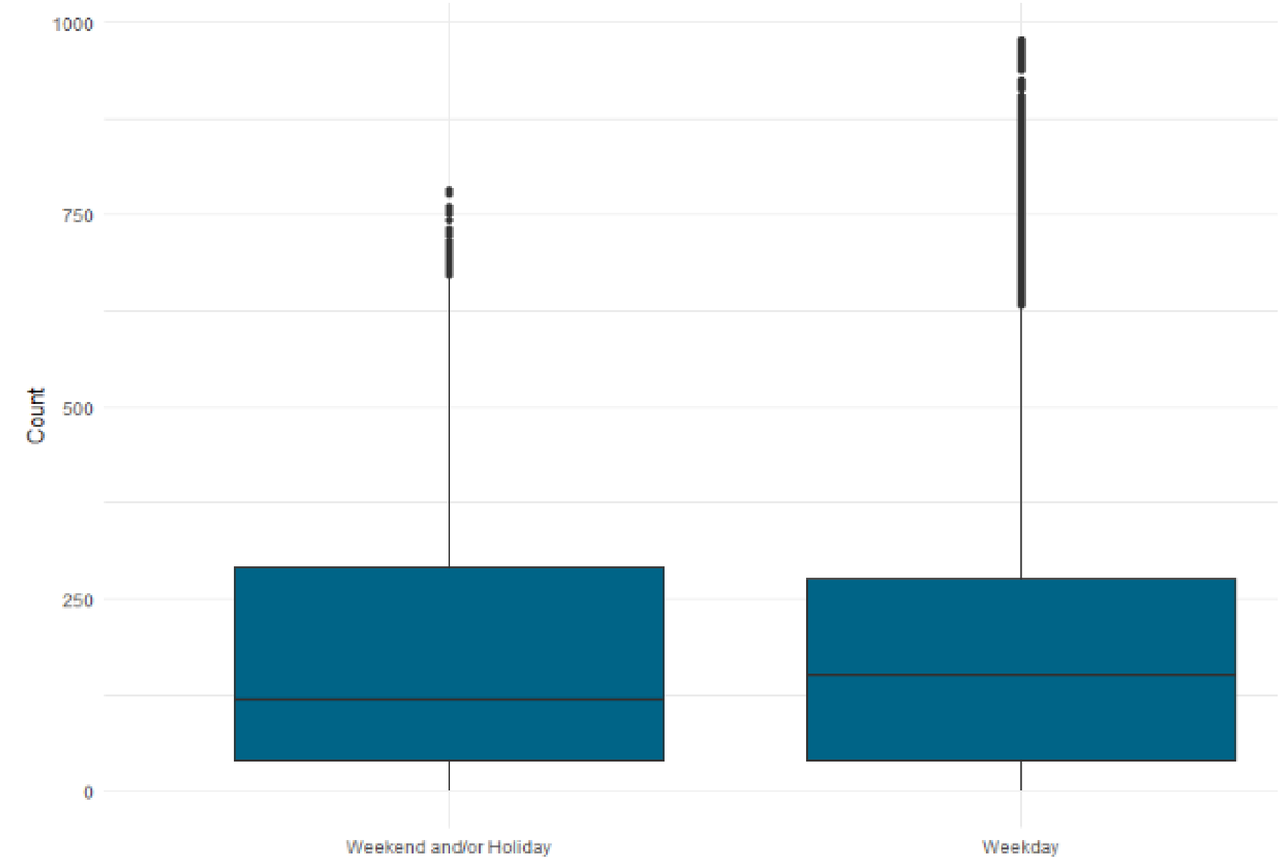


EDA

Hypothesis

Rental demand increases under the following conditions:

- Warmer temperatures and favorable weather (e.g., clear or partly cloudy conditions).
- Weekends and holidays, due to increased leisure activity.
- Weekday mornings and afternoons, reflecting commuter patterns.

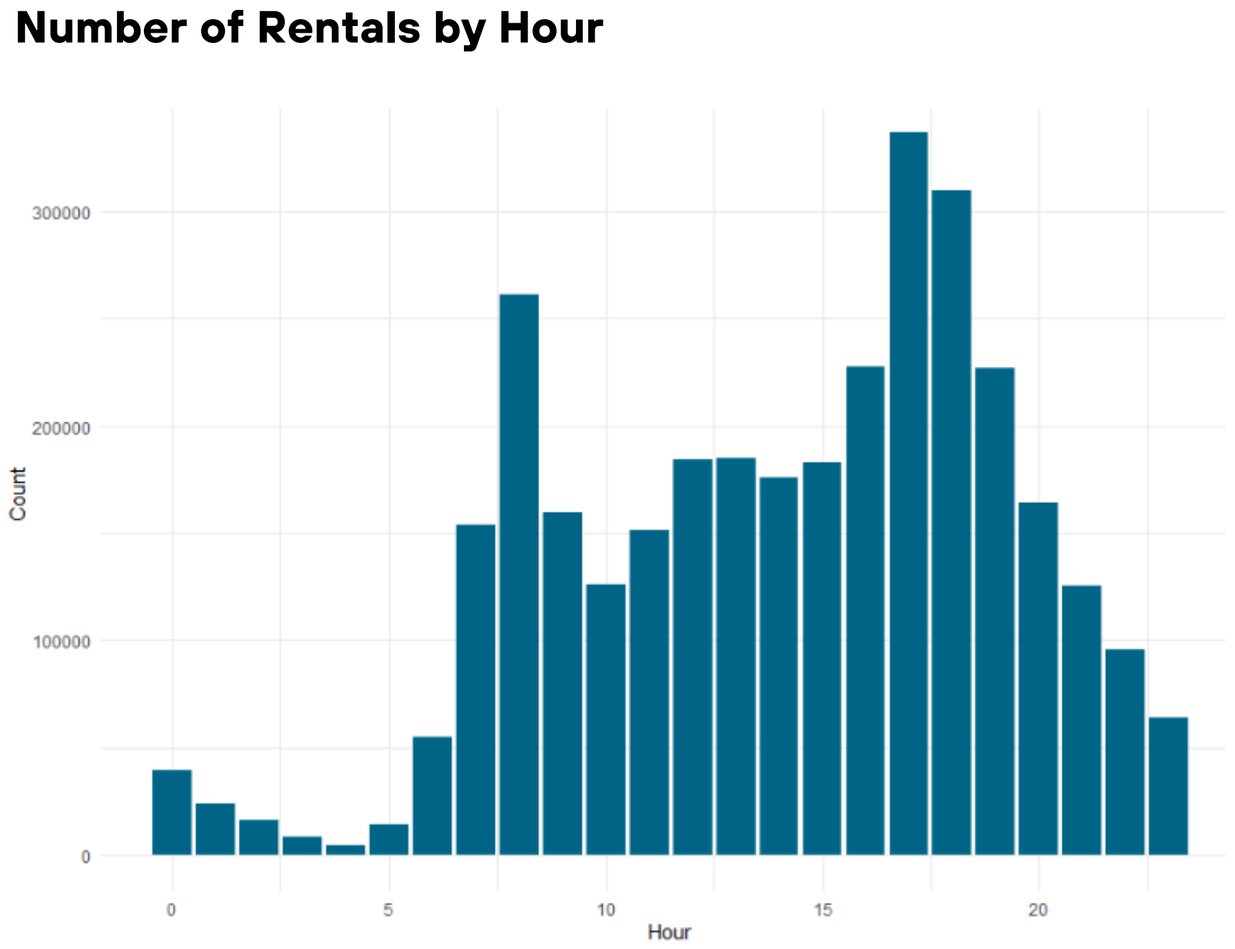


EDA

Hypothesis

Rental demand increases under the following conditions:

- Warmer temperatures and favorable weather (e.g., clear or partly cloudy conditions).
- Weekends and holidays, due to increased leisure activity.
- Weekday mornings and afternoons, reflecting commuter patterns.



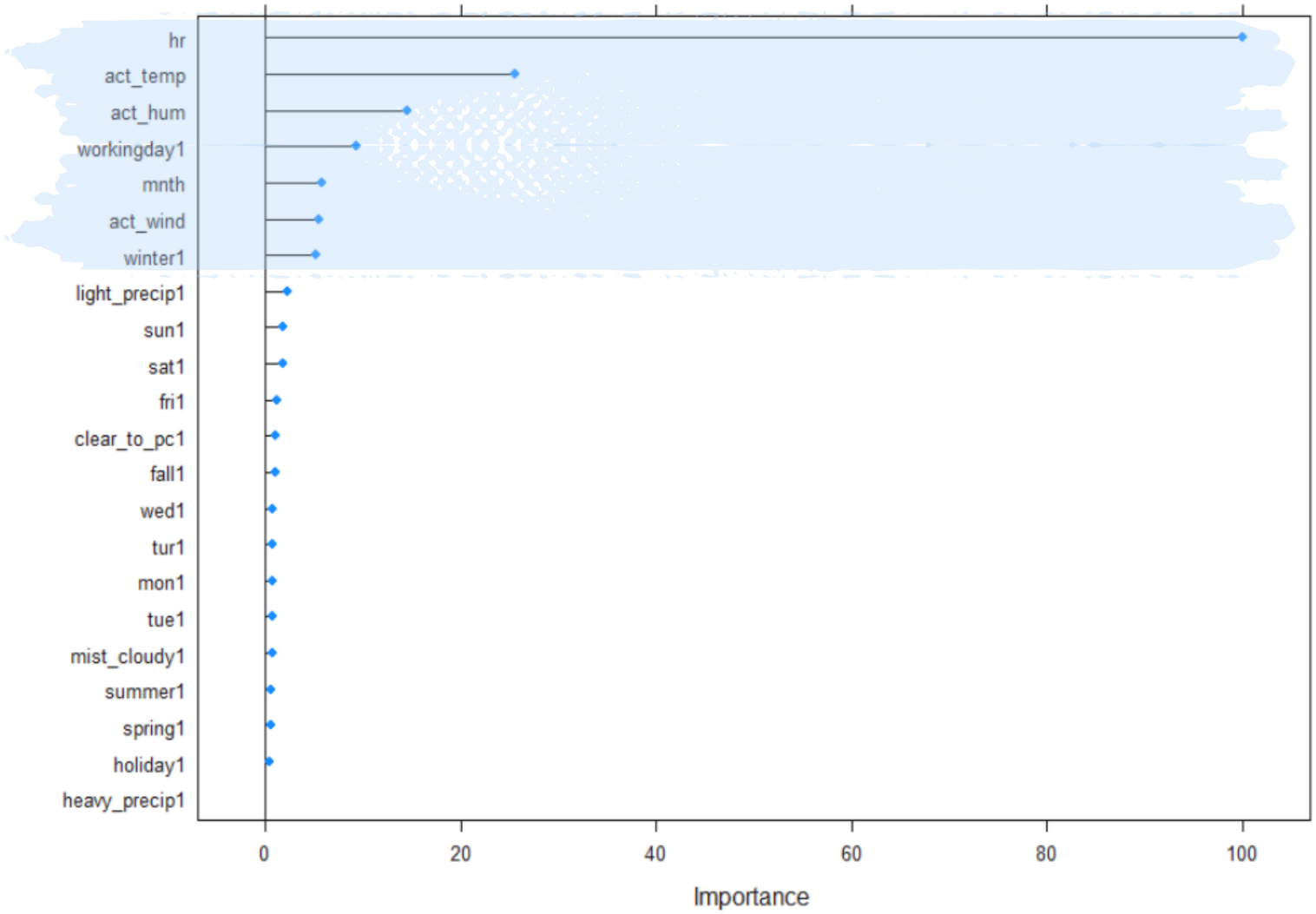
Model Results

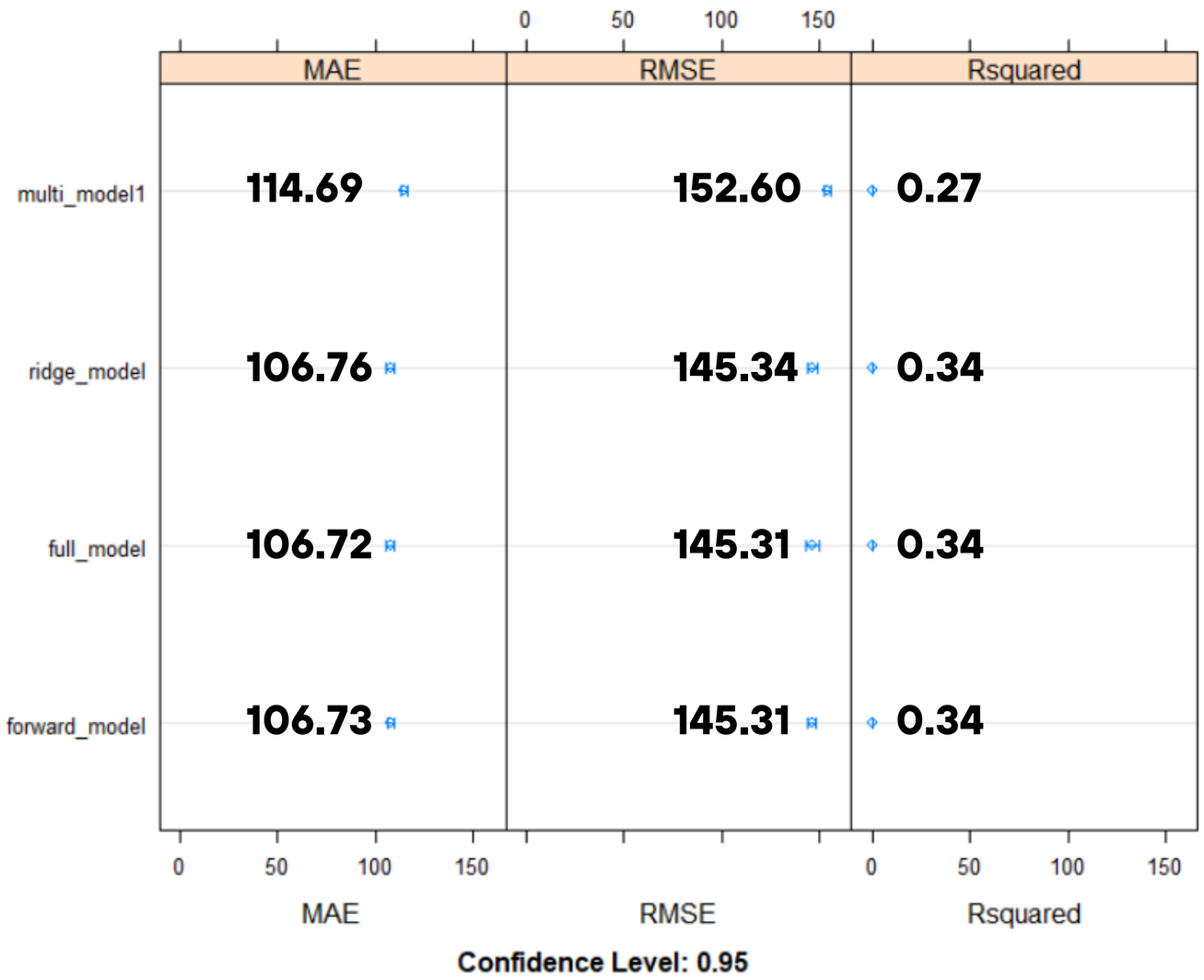
Key Variables

Certain variables consistently contributed to predictive performance across multiple models.

Full feature models outperformed models using only a subset of top-ranked variables.

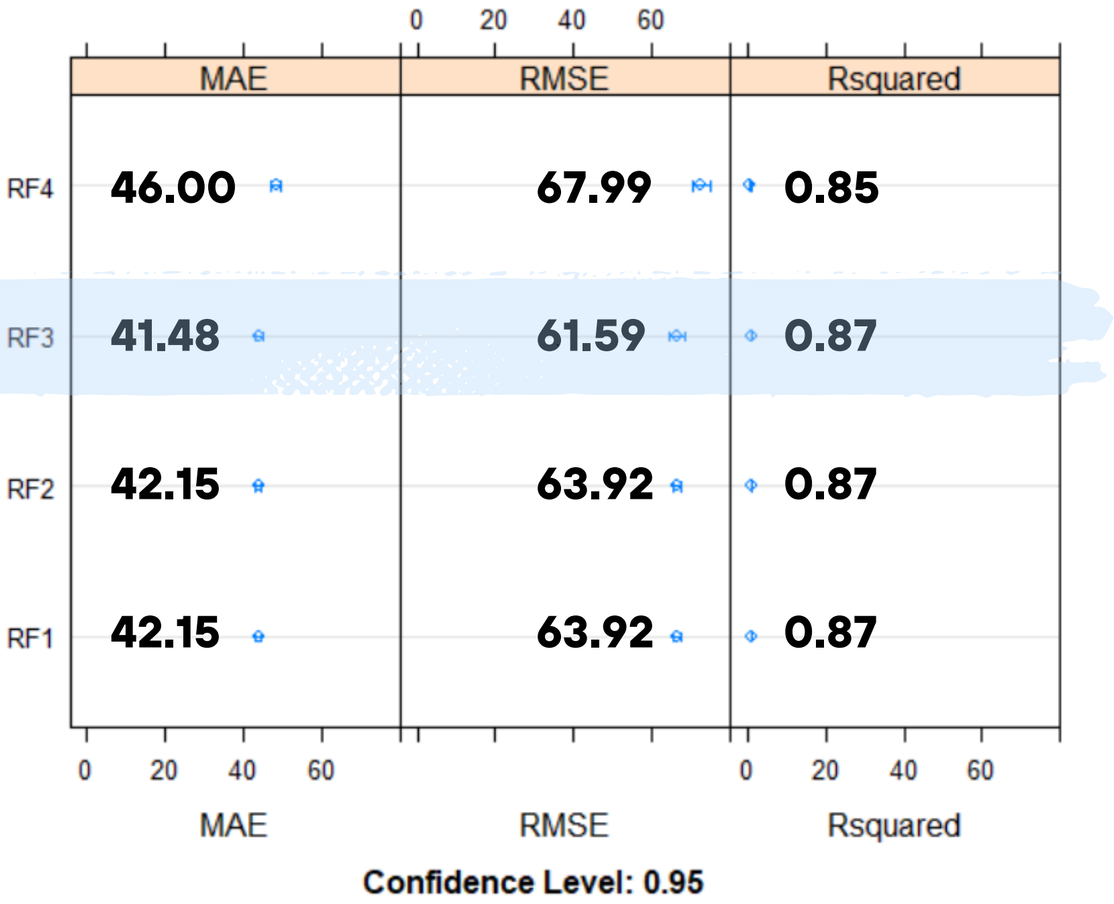
This performance gap is likely due to the exclusion of important seasonal and weather-related features in the reduced models, despite their significance being clearly indicated during the exploratory analysis.





Linear Model Comparisons

- Full models for each—except multi model, multi includes seasons, weather, temperature, and humidity.
- Normalized values.
- Model matrix columns.



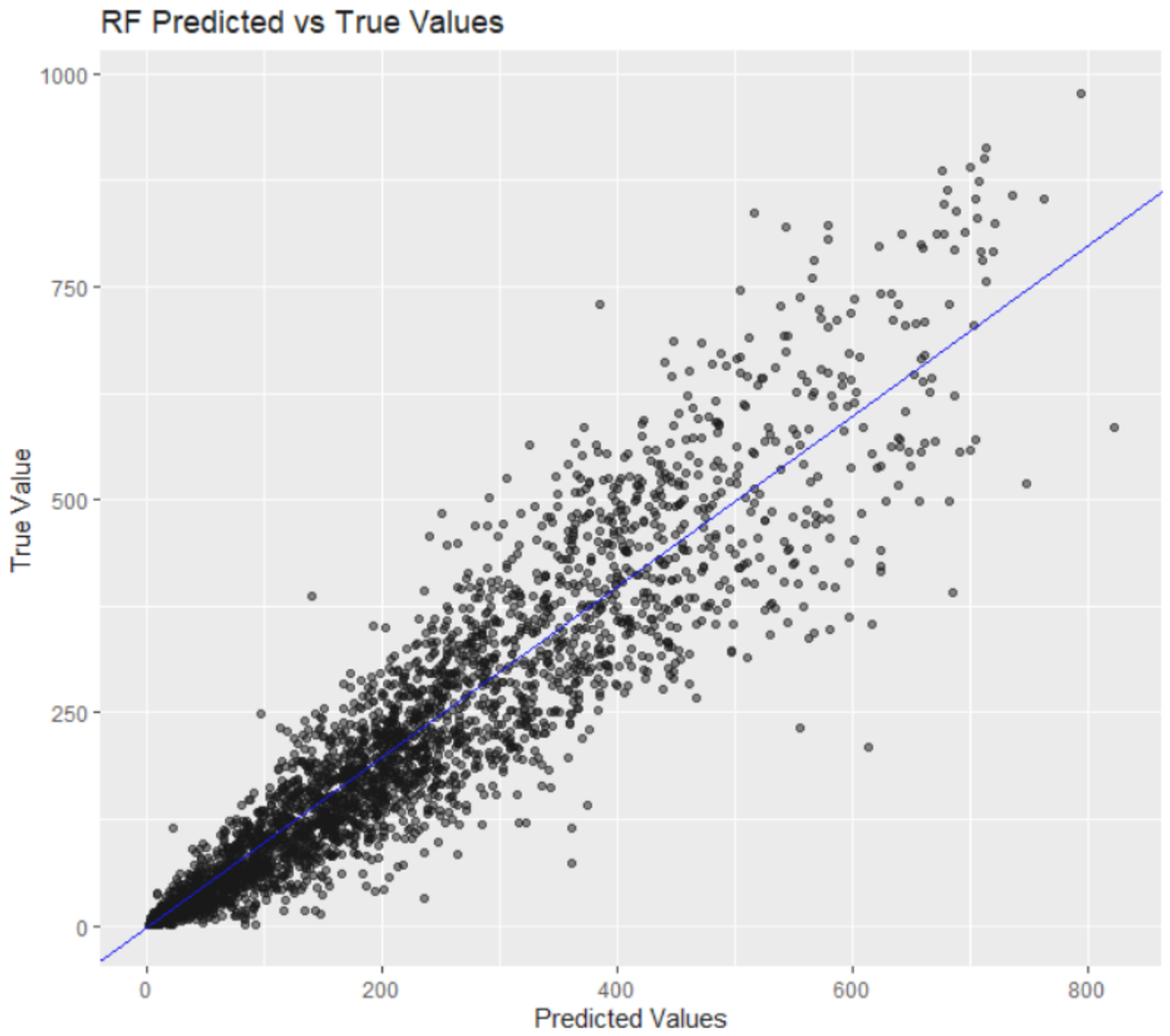
Random Forest Model Comparisons

- Started with full models for each—tree, bagged tree, random forest, and XGboost.
- Tried models with normalized and denormalized values.
- Tested models with and without model matrix (binary) columns.
- Filtered from the best full models and tested them on reduced number of features.
- All configurations for XGboost models RMSE = 70.

Final Model Results

Error Plot

Diverges more with the higher counts.

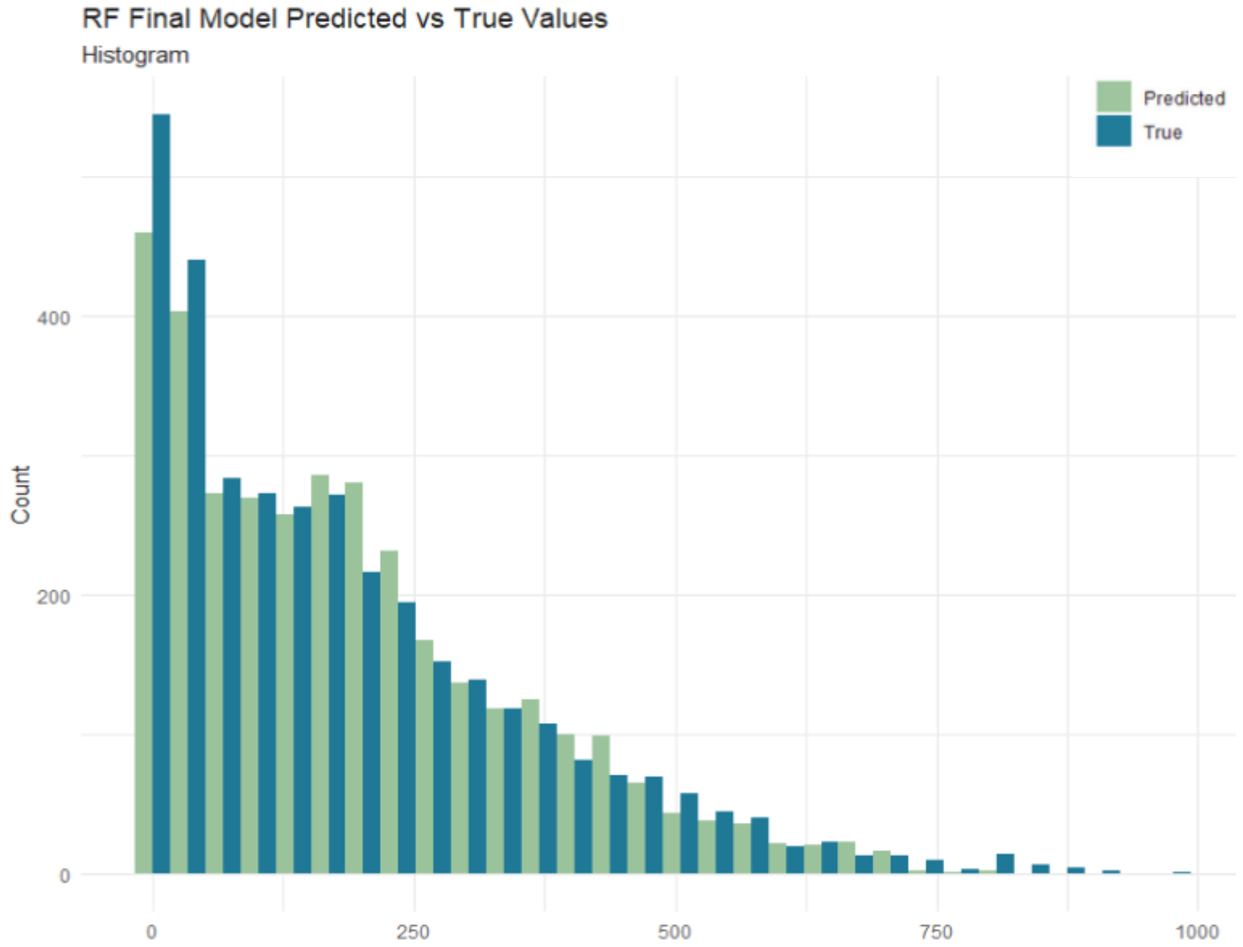


Final Model Results

Predictions lower than the actual counts at higher numbers of rentals.

Predictions start becoming closer to actual counts with more average rental counts.

Predictions become much lower with low true rentals.



Conclusion

Initial Overfitting

Early models showed signs of overfitting, largely due to the inclusion of casual and registered user counts. Such variables that directly sum to the target and introduce data leakage.

Normalization Confusion

We initially didn't realize that temperature, humidity, and wind speed were already normalized.

After further research, we converted these back to more interpretable values to better understand their relationship to rental behavior.

Feature Selection Considerations

While full-feature models performed well, they may not always be ideal in practice. However, many of the included features like weather and day of the week, are commonly available and easy to collect.

Next Steps

Adding a margin of error or confidence interval would help to better guide resource planning.

Experiment with alternative model configurations like feature subsets and regularization.

Test additional model types to explore performance improvements.

Consider incorporating external data such as local events and traffic to enhance predictive accuracy.

Code Appendix

```
# Ranger Method faster than RF

my_control <- trainControl(method = 'cv',
      number = 5)

rf_model <- train(cnt ~ .,
                  bike_train,
                  method = 'ranger',
                  ntree = 250,
                  trControl = my_control)
```

```
# Un-normaliz the weather data

act_temp <- bike_new$temp*41
act_wind <- bike_new$windspeed*67
act_hum <- bike_new$hum*100

bike_act <- cbind(bike_new, act_temp, act_wind, act_hum)
bike_act <- select (bike_act, -temp, -atemp, -windspeed, -hum)

head(bike_act)
```

```
# True vs Predictions Histogram

true_values <- bike_test_act$cnt
pred_values <- rf_3

true_vs_pred <- data.frame(true = true_values,
      predicted = pred_values)

predictions <- pivot_longer(true_vs_pred, c(true,
      predicted), names_to = "type", values_to = "count")
```